NLP Techniques and Text Retrieval

Kenneth W. Church
AT&T Bell Laboratories

Classical Information Retrieval (IR)

- Sort a collection of documents
  (e.g., scientific papers, abstracts, paragraphs)
  by how much they ‘‘match’’ a query

- The query can be a sequence of keywords
  or arbitrary text
  (e.g., one of the documents)

Motivation

- Text is available like never before
  - Currently, $N \approx 100$ million words,
  - and projections run as high as 10^{15} bytes by 2000!

- What can we do with it all?

- It is better to do something simple,
  than nothing at all.

- IR vs. Natural Language Understanding

- Revival of 1950-style empiricism

Pattern Recognition Problems in CL

- Information Retrieval:
  - Sort docs by how much they match a query.

- Author Identification:
  - Is this doc more like author A's docs or author B's docs?

- Content Analysis:
  - Is the context of this use of *bank* more like sense 1's contexts or like sense 2's contexts?

- Machine Translation
  - Is the context of this use of *drug* more like those that were translated as *drogue* or those that were translated as *medicament*?

- Keywords (and boolean combinations thereof)

- Vector-Space ‘‘Model’’ (Salton, chap 10.1)

  - Represent the query and the documents as V-dimensional vectors

  - sort vectors by cosine distance

$$sim(x, y) \quad = \quad cos(x, y) \quad = \quad \frac{\sum_{i=1}^{V} x_i \, y_i}{|x| \; |y|}$$

$$|x|^2 \; = \; \sum_{i=1}^{V} x_i^2$$

- Probabilistic Retrieval Model (Salton, chap 10.3)

  Sort documents *d* by

$$score(d) \; = \; \prod_{token \text{ in } d} \frac{Pr(token|rel)}{Pr(token|irrel)}$$

Commerical Products

- There are a number of products
- mostly, keyword systems
- except DowQuest (by Dow Jones & TMC)

- Few are profitable, except Lexis (MDC), which is effectively subsidized by the U.S. government.

- Since libraries are underfunded (and under-appreciated), it is hard to make a good business case for an IR start-up.

- IR is currently very expensive
  (distribution + archival costs)

- Brightspot: CDROMs (PC Magazine, OED)
  - distribution costs: less than telephone
  - archival costs: less than microfiche

Bellcore's Example

- Human-Computer Interaction

  c1:Human machine interface for Lab ABC computer applications

  c2:A survey of user opinion of computer system response time

  c3:The EPS user interface management system

  c4:System and human system engineering testing of EPS

  c5:Relation of user-perceived response time to error measurement

- Graphs

  m1:The generation of random, binary, unordered trees

  m2:The intersection graph of paths in trees

  m3:Graph minors IV: Widths of trees and well-quasi-ordering

  m4:Graph minors: A survey

Term-by-Document Matrix

c1	c2	c3	c4	c5	m1	m2	m3	m4	
1			1						human
1		1							interface
1	1								computer
	1	1		1					user
	1	1	2						system
	1			1					response
	1			1					time
		1	1						EPS
	1							1	survey
					1	1	1		trees
						1	1	1	graph
							1	1	minors

- By eye, it looks like the documents separate into two groups (c*/ m*), and

- the terms separate into two groups (above/ below *survey*).

- The various automatic procedures will hopefully come to the same conclusion.

Document Cosines

$$sim(x, y) \quad = \quad cos(x, y) \quad = \quad \frac{\sum_{i=1}^{V} x_i \, y_i}{|x| \; |y|}$$

$$cos(c1, c2) \quad = \quad \frac{1}{\sqrt{3} \times \sqrt{6}} \quad = \quad 0.2$$

	c1	c2	c3	c4	c5	m1	m2	m3	m4
c1	1.0	0.2	0.3	0.2					
c2	0.2	1.0	0.4	0.3	0.7				0.2
c3	0.3	0.4	1.0	0.6	0.3				
c4	0.2	0.3	0.6	1.0					
c5		0.7	0.3		1.0				
m1						1.0	0.7	0.6	
m2						0.7	1.0	0.8	0.4
m3						0.6	0.8	1.0	0.7
m4		0.2					0.4	0.7	1.0

Term Cosines

human	interface	computer	user	system	response	time	EPS	survey	trees	graph	minors	
1.0	0.5	0.5		0.6			0.5					human
0.5	1.0	0.5	0.4	0.3			0.5					interface
0.5	0.5	1.0	0.4	0.3	0.5	0.5		0.5				computer
	0.4	0.4	1.0	0.5	0.8	0.8	0.4	0.4				user
0.6	0.3	0.3	0.5	1.0	0.3	0.3	0.9	0.3				system
		0.5	0.8	0.3	1.0	1.0		0.5				response
		0.5	0.8	0.3	1.0	1.0		0.5				time
0.5	0.5		0.4	0.9			1.0					EPS
		0.5	0.4	0.3	0.5	0.5		1.0		0.4	0.5	survey
									1.0	0.7	0.4	trees
								0.4	0.7	1.0	0.8	graph
								0.5	0.4	0.8	1.0	minors

Hierarchical Clustering (hclust)

- Useful display for eye-balling a similarity matrix, e.g., cos of terms, cos of docs

- Can also play a role in IR

    - Term clusters can be used to reduce dimensionality (V $\rightarrow$ number of clusters)

    - Doc clusters can be used to reduce search space (score clusters, rather than documents)

- Input: a similiarity matrix (square, symmetric)

- Output: a tree of clusters

- Methods: single linkage, complete linkage, k-means, and many more

- I'm using the implementation in S (Becker *et al.*), with the default method (complete linkage).

Hclust of Terms

Hclust of Docs

Correlations

- It is generally assumed that terms are indep.
  That is, $\rho_{i,j} = 0$ when $i \neq j$

- In practice, this assumption is often problematic.

- Positive correlations arise when two words share similar distributions:
  - synonymous terms: *computer*, *machine*
  - morphological variants: *computer*, *computers*
  - spelling variants: *IBM*, *I.B.M.*
  - upper and lower case: *computer*, *Computer*
  - strong collocations: *computer scientist*

- Negative correlations arise when two words have complementary distributions.

- Note the large correlations in this example.

human	interface	computer	user	system	response	time	EPS	survey	trees	graph	minors	
1.0	0.4	0.4	–0.4	0.4	–0.3	–0.3	0.4	–0.3	–0.4	–0.4	–0.3	human
0.4	1.0	0.4	0.2	0.04	–0.3	–0.3	0.4	–0.3	–0.4	–0.4	–0.3	interface
0.4	0.4	1.0	0.2	0.04	0.4	0.4	–0.3	0.4	–0.4	–0.4	–0.3	computer
–0.4	0.2	0.2	1.0	0.2	0.8	0.8	0.2	0.2	–0.5	–0.5	–0.4	user
0.4	0.04	0.04	0.2	1.0	0.04	0.04	0.8	0.04	–0.5	–0.5	–0.3	system
–0.3	–0.3	0.4	0.8	0.04	1.0	1.0	–0.3	0.4	–0.4	–0.4	–0.3	response
–0.3	–0.3	0.4	0.8	0.04	1.0	1.0	–0.3	0.4	–0.4	–0.4	–0.3	time
0.4	0.4	–0.3	0.2	0.8	–0.3	–0.3	1.0	–0.3	–0.4	–0.4	–0.3	EPS
–0.3	–0.3	0.4	0.2	0.04	0.4	0.4	–0.3	1.0	–0.4	0.2	0.4	survey
–0.4	–0.4	–0.4	–0.5	–0.5	–0.4	–0.4	–0.4	–0.4	1.0	0.5	0.2	trees
–0.4	–0.4	–0.4	–0.5	–0.5	–0.4	–0.4	–0.4	0.2	0.5	1.0	0.8	graph
–0.3	–0.3	–0.3	–0.4	–0.4	–0.3	–0.3	–0.3	0.4	0.2	0.8	1.0	minors

Hclust of Correlation Coefs

Methods of Correcting for Correlations

- One can compute the correlation for each pair of terms, and adjust the cos calculation appropriately.

- Unfortunately, this is generally not practical since there are V^2 correlations to consider.

- For any particular pair of documents, one can look at the terms that contribute the most and adjust for their correlations. (I don't think this has been tried.)

- It is also quite common to merge terms that have ''similar'' distributions, either for linguistic or statistical reasons.

- For example, it is common to treat morphologically related words (e.g., *computer* and *computers*) as a single term.

- Treating two words (*i* and *j*) as the same term is equivalent to assuming $\rho_{i,j} \approx 1$.

‘‘Automatic’’ Thesaurus

- Merge terms that cluster together
  - human/ interface/ computer
  - user/ response/ time
  - system/ EPS
  - graph/ minors

Before

c1	c2	c3	c4	c5	m1	m2	m3	m4	
1			1						human
1		1							interface
1	1								computer
	1	1		1					user
	1	1	2						system
	1			1					response
	1			1					time
		1	1						EPS
	1							1	survey
					1	1	1		trees
						1	1	1	graph
							1	1	minors

After

c1	c2	c3	c4	c5	m1	m2	m3	m4	
3	1	1	1						com/hum/inter
	3	1		3					user/res/time
	1	2	3						system/EPS
	1							1	survey
					1	1	1		trees
						1	2	2	graph/minors

Cosines after Term Reduction

	c1	c2	c3	c4	c5	m1	m2	m3	m4
c1	1.0	0.3	0.4	0.3					
c2	0.3	1.0	0.7	0.4	0.9				0.1
c3	0.4	0.7	1.0	0.9	0.4				
c4	0.3	0.4	0.9	1.0					
c5		0.9	0.4		1.0				
m1						1.0	0.7	0.4	
m1						0.7	1.0	0.9	0.6
m3						0.4	0.9	1.0	0.8
m4		0.1					0.6	0.8	1.0

Differences (before – after)

	c1	c2	c3	c4	c5	m1	m2	m3	m4
c1		–0.05	–0.1	–0.08					
c2	–0.05		–0.3	–0.03	–0.2				0.1
c3	–0.1	–0.3		–0.3	–0.1				
c4	–0.08	–0.03	–0.3						
c5		–0.2	–0.1						
m1								0.1	
m1								–0.1	–0.2
m3						0.1	–0.1		–0.1
m4		0.1					–0.2	–0.1	

Singular Value Decomposition (SVD)

- Merging terms (either for linguistic or statistical reasons) is equivalent to assuming $\rho \approx 1$.

- Related terms (e.g., *computer* and *human*) often have moderate correlations (e.g. 0.4), much greater than 0 (total independence), but well below 1 (total dependence).

- Bellcore has proposed SVD analysis as a way of transforming the term-by-document matrix into an orthogonal space of smaller dimensionality.

- They have found that SVD analysis improves performance, probably because it removes troublesome correlations.

- In addition, SVD acts as a smoothing process, and may actually predict the truth better than the raw data.

- Let X be the term-by-doc matrix described above

- X can be decomposed into the product of three matrices, such that:

  - $X \approx T_0 \; S_0 \; D_0'$

  - T_0 and D_0 have orthogonal unit length cols so that $T_0 \; T_0' \approx I$ and $D_0 \; D_0' \approx I$

  - S_0 is a diagonal matrix.

T0									
human	0.22	–0.11	0.29	–0.41	–0.11	–0.34	0.52	–0.06	–0.41
interface	0.20	–0.07	0.14	–0.55	0.28	0.50	–0.07	–0.01	–0.11
computer	0.24	0.04	–0.16	–0.59	–0.11	–0.25	–0.30	0.06	0.49
user	0.40	0.06	–0.34	0.10	0.33	0.38			0.01
system	0.64	–0.17	0.36	0.33	–0.16	–0.21	–0.17	0.03	0.27
response	0.27	0.11	–0.43	0.07	0.08	–0.17	0.28	–0.02	–0.05
time	0.27	0.11	–0.43	0.07	0.08	–0.17	0.28	–0.02	–0.05
EPS	0.30	–0.14	0.33	0.19	0.11	0.27	0.03	–0.02	–0.17
survey	0.21	0.27	–0.18	–0.03	–0.54	0.08	–0.47	–0.04	–0.58
trees	0.01	0.49	0.23	0.03	0.59	–0.39	–0.29	0.25	–0.23
graph	0.04	0.62	0.22		–0.07	0.11	0.16	–0.68	0.23
minors	0.03	0.45	0.14	–0.01	–0.30	0.28	0.34	0.68	0.18

S0	
	3.34
	2.54
	2.35
	1.64
	1.5
	1.31
	0.85

0.56

0.36

D0									
c1	0.20	–0.06	0.11	–0.95	0.05	0.08	0.11	–0.01	–0.06
c2	0.61	0.17	–0.50	–0.03	–0.21	–0.26	–0.43	0.05	0.24
c3	0.46	–0.13	0.21	0.04	0.38	0.72	–0.24	0.01	0.02
c4	0.54	–0.23	0.57	0.27	–0.21	–0.37	0.26	–0.02	–0.08
c5	0.28	0.11	–0.51	0.15	0.33	0.03	0.67	–0.06	–0.26
c6		0.19	0.10	0.02	0.39	–0.30	–0.34	0.45	–0.62
c7	0.01	0.44	0.19	0.02	0.35	–0.21	–0.15	–0.76	0.02
c8	0.02	0.62	0.25	0.01	0.15		0.25	0.45	0.52
c9	0.08	0.53	0.08	–0.03	–0.60	0.36	0.04	–0.07	–0.45

- We now truncate all three matrixes to just the first few cols

$$X \approx \hat{X} = T\ S\ D'$$

- The claim is that the first few cols capture most of the variance, and that the other cols are mostly noise due to random variations in the training data.

	T		S		D		
human	0.22	–0.11	3.34	0	c1	0.20	–0.06
interface	0.20	-0.07	0	2.54	c2	0.61	0.17
computer	0.24	0.04			c3	0.46	–0.13
user	0.40	0.06			c4	0.54	–0.23
system	0.64	–0.17			c5	0.28	0.11
response	0.27	0.11			c6		0.19
time	0.27	0.11			c7	0.01	0.44
EPS	0.30	–0.14			c8	0.02	0.62
survey	0.21	0.27			c9	0.08	0.53
trees	0.01	0.49					
graph	0.04	0.62					
minors	0.03	0.45					

Cosines after SVD

- Reconstitute $X \approx \hat{X} = T\ S\ D'$, and compute cosines for all pairs of documents.

- Note that the cosines now do a better job of separating the c docs from the m docs.

	c1	c2	c3	c4	c5	m1	m2	m3	m4
c1	1.0	0.9	1.0	1.0	0.9	–0.2	–0.2	–0.2	–0.02
c2	0.9	1.0	0.9	0.9	1.0	0.2	0.2	0.3	0.4
c3	1.0	0.9	1.0	1.0	0.9	–0.2	–0.2	–0.2	–0.01
c4	1.0	0.9	1.0	1.0	0.8	–0.3	–0.3	–0.3	–0.1
c5	0.9	1.0	0.9	0.8	1.0	0.3	0.3	0.3	0.5
m1	–0.2	0.2	–0.2	–0.3	0.3	1.0	1.0	1.0	1.0
m2	–0.2	0.2	–0.2	–0.3	0.3	1.0	1.0	1.0	1.0
m3	–0.2	0.3	–0.2	–0.3	0.3	1.0	1.0	1.0	1.0
m4	–0.02	0.4	–0.01	–0.1	0.5	1.0	1.0	1.0	1.0

Hclust of Document Cosines after SVD

SVD terms

SVD docs

Evaluation

$$\text{precision} \equiv \frac{\text{number of relevant \& retrieved documents}}{\text{number of retrieved documents}}$$

$$\text{recall} \equiv \frac{\text{number of relevant \& retrieved documents}}{\text{number of relevant documents}}$$

Standarized Datasets (Ed Fox's CDROM):

- MED, CACM, ADI, CISI, CRAN, TIME
- Queries, Documents, Relevance Judgments

Private Datasets:

- Bellcore Memos (*Who Knows*)
- Associated Press Newswire

Bellcore Figure 4

Term Weighting & Stop-list

- Emphasize content words and de-emphasize function words

$$sim(x, y) = \frac{\sum_{t=1}^{V} (w_t\ x_t)\ (w_t\ y_t)}{|w\ x|\ |w\ y|}$$

- IDF (inverse document freq)

$$w_t = -\log_2 \frac{\text{number of documents with term } t}{\text{number of documents (= N)}}$$

- Entropy/ Noise Weighting

$$p_{t,d} = \frac{f_{t,d}}{\sum_{i=1}^{N} f_{t,d}}$$

$$H_t = -\sum_{i=1}^{N} p_{t,i} \log_2 p_{t,i}$$

Appendix to Harman’s paper I

Appendix to Harman’s paper II

Readings for Next Time

- Salton's Experience with my Part of Speech Tagger for applications in IR
- How Effective is Suffixing?
- Attempts to use Syntax for IR at UMass
- Bilingual IR

Summary of Lecture I

- IR Problem: sort docs by $sim(d,q)$

- Vector-space ‘‘Model’’ (cosine similarity)
- Probabilistic Retrieval Model

- Clustering

- Correlations ‘‘Fixes’’
  - Merge morphologically related words
  - Merge synonymous words using a thesaurus
  - Singular Value Decomposition (SVD)

- Evaulation

- Term Weighting: IDF & Entropy

Lecture II

Readings

1. Salton's Experience with my Part of Speech Tagger for applications in IR

2. How Effective is Suffixing?

3. Attempts to use Syntax for IR at UMass

4. Bilingual IR

- I tried to pick papers that would be encouraging (e.g., 4).

- Much of the IR literature can be depressing, especially to those of us working in NLP.

- If something as basic as suffixing is of marginal utility, then maybe we're in the wrong business.

Lesson: Get Longer Queries

- Single best improvement

- You can work very hard on
  improving the statistics, text normalization, etc.,
  but nothing increases precision/ recall
  like LONG queries with lots of bits

- Wanted: methods to help users generate longer queries
  a. User Profiles
  b. Relevance Feedback

- Users tend to generate short queries with very few bits

Queries from MED dataset

- the crystalline lens in vertebrates, including humans.

- the relationship of blood and cerebrospinal fluid oxygen concentrations or partial pressures. a method of interest is polarography.

- electron microscopy of lung or bronchi.

- tissue culture of lung or bronchial neoplasms.

Queries from TIME dataset

- KENNEDY ADMINISTRATION PRESSURE ON NGO DINH DIEM TO STOP SUPPRESSING THE BUDDHISTS .

- EFFORTS OF AMBASSADOR HENRY CABOT LODGE TO GET VIET NAM'S PRESIDENT DIEM TO CHANGE HIS POLICIES OF POLITICAL REPRESSION .

- NUMBER OF TROOPS THE UNITED STATES HAS STATIONED IN SOUTH VIET NAM AS COMPARED WITH THE NUMBER OF TROOPS IT HAS STATIONED IN WEST GERMANY .

- U.S . POLICY TOWARD THE NEW REGIME IN SOUTH VIET NAM WHICH OVERTHREW PRESIDENT DIEM .

- PERSONS INVOLVED IN THE VIET NAM COUP .

Relevance Feedback

- Documents tend to be much longer than queries
- (because authors spend more time composing documents than users spend composing queries)
- Don’t ask users to compose a query
- Rather, get them to give you a set of relevant documents and use that as the query
- Relevance Feedback = Call-by-plagiarism

Vector-Space Search on ‘‘Obit-White’’ Story

cos	date	time	headword	title
1.00	04-09-90	10:45	Obit-White	Ryan White Taught Nation to Care About AIDS, Friends Say
0.48	04-08-90	18:08	White-Reax	Americans Pay Tribute to AIDS Victim Ryan White
0.45	04-10-90	03:11	RyanWhite	AIDS Victim Ryan White Remembered as Teacher, Student
0.41	04-07-90	19:36	Ryan’sLegacy	Unassuming Indiana Teen-ager Taught America About AIDS
0.41	04-08-90	12:08	Bush-White	Bush Mourns Boy’s Death
0.39	04-11-90	21:42	RyanWhite	1,500 Say Goodbye to AIDS Victim Ryan White
0.36	04-03-90	14:03	RyanWhite	Ryan White, AIDS Patient, Near Death; Well-Wishers’ Calls Flood H
0.32	04-02-90	21:32	RyanWhite	AIDS Victim Unconscious, Said to Be Dying from Internal Bleeding
0.31	04-02-90	15:09	RyanWhite	AIDS Patient Ryan White Reported Near Death
0.29	04-12-90	11:23	RyanWhite	First Lady, Celebrities Attend Funeral for Young AIDS Victim
0.29	04-03-90	04:18	RyanWhiteChrono	
0.28	04-04-90	13:35	RyanWhite	AIDS Patient Ryan White Remains in Critical Condition
0.28	04-04-90	22:37	RyanWhite	Town That Once Spurned Ryan White Joins Nation in Wishing Him
0.28	04-09-90	02:47	White-Chronolog	White’s Struggle With AIDS Began At Age 13
0.26	04-11-90	12:35	RyanWhite	Barbara Bush, Celebrities to Attend Ryan White Funeral
0.26	04-25-90	20:23	RyanWhite	Victim’s Mother Lobbies for AIDS Bill
0.25	04-26-90	03:20	RyanWhite	AIDS Victim’s Mother Lobbies for Spending Bill
0.25	04-03-90	19:54	RyanWhite-Chron	
0.23	04-05-90	02:56	RyanWhite	In City That Barred AIDS Boy, an Outpouring of Sympathy
0.22	04-10-90	23:05	RyanWhite	Hundreds Pay Respect at Funeral Home
0.21	04-09-90	01:49	WashingtoninBri	Lithuanian Declaration Should Be Withdrawn, Soviet Says
0.17	04-06-90	08:01	RyanWhite	Ryan White Brings Hope to Other Patients with Gifts from Elton Joh
0.16	04-09-90	04:36	FarmAid	Benefit Urges Fans to Message Congress
0.16	04-03-90	13:48	Bush	Bush Plants Elm Tree Dedicated to Ryan White
0.15	04-02-90	19:50	Ryan-S&Ls	Republicans Ask for Speedy Vote on S&L Nominee
0.15	04-09-90	06:38	Quotes	
0.15	04-09-90	04:50	DigestBriefs	
0.14	04-03-90	02:11	RyanNomination	Republicans Optimistic That S&L Nominee Will Be Confirmed
0.14	04-08-90	02:19	FarmAid	Farm Aid Benefit Concert Returns With Environmental Focus
0.14	04-05-90	09:24	S&LDirector	Ryan Wins Surprisingly Easy Conformation
0.14	04-03-90	18:15	Bush	Bush Plants Tree to Symbolize Clean Air; Dedicates It to Ryan White
0.14	04-20-90	19:32	AIDSBoy	Church Bars Boy With AIDS, Then Reverses Itself
0.14	04-19-90	04:01	AIDSBrothers	One of Three Hemophiliac Brothers Gets AIDS
0.13	04-18-90	02:43	AIDSSpending	Survey Says Experts Split on Diversion of Funds for AIDS Research
0.13	04-11-90	20:50	DigestBriefs/1	
0.12	04-16-90	11:56	RuralAIDS	AIDS Panel Studies Special Problems Of Rural Patients

Salton's Experience with my Parts Program

- Goal: an automatic procedure for identifying phrases like *computer scientist*.

- Proposal: filter the output of my program with a condition based on IDF weighting.

- Results look promising, as demonstrated by the appendix.

- However, the results may not produce very large improvements in precision/ recall.

- At least, Lewis & Croft didn't find very much improvement (though their study differs in a number of ways, so it may not be comparable).

- Side remark: IR demands simple and robust methods

- PLNLP is apparently less useful to Salton because it is too ambitious.

My Parts Program

- He will table/ VB the motion
- The table/ NN is ready

Applications

- Speech Synthesis
- Speech Recognition
- Optical Character Recognition (OCR)
- Spelling Correction
- Information Retrieval (IR)
- Proof-Reading (WWB)
- Query Answering (Q&A)
- Machine Translation (MT)
- Tagging Corpora for future research (COBUILD)

Speech Synthesis Applications

- The *wind* is strong.
- Don’t forget to *wind* your watch.

- Did you see *THAT* ?
- It is a shame *that* he’s leaving.

- oily FLUID
- TRANSMISSION fluid

Proposed Method

Linear time dynamic programming algorithm to find the assignment of parts of speech to words that optimizes the product of

- Lexical Probabilities: Prob of part of speech given word

- Contextual Probabilities: Prob of part of speech given context (next two parts of speech)

Probability estimates were obtained by training on the Tagged Brown Corpus

It Works!

- Performance is very encouraging

- The program has been run on a year of the Wall Street Journal and several years of the AP News Wire (hundreds of millions of words)

- Reliable enough to support the needs of a growing user population

Lexical Probabilities of Some Common Words

Word	Cat	P(Cat | Word)	Cat	P(Cat | Word)
costs	VBZ	0.08	NNS	0.92
human	NN	0.12	JJ	0.88
humans	NNS	1.00		
amount	VB	0.18	NN	0.82
amounts	VBZ	0.46	NNS	0.54
meeting	VBG	0.20	NN	0.80
attack	VB	0.25	NN	0.75
mark	VB	0.30	NN	0.70
support	VB	0.31	NN	0.69
change	VB	0.32	NN	0.68
sacrifice	VB	0.37	NN	0.63
use	VB	0.38	NN	0.62
gathering	VBG	0.45	NN	0.55
strike	VB	0.46	NN	0.54
sink	VB	0.48	NN	0.52
travel	VB	0.49	NN	0.51

Typical Contextual Probabilities

log P	(cat | __ cat, cat)			log P	(cat | __ cat, cat)		
–0.81	IN	AT	NNS	–4.31	VBN	AT	NN
–0.83	IN	AT	NN	–4.47	VBN	AT	NNS
–2.28	.	AT	NNS	–4.49	NN	AT	NN
–2.51	.	AT	NN	–4.67	NN	AT	NNS
–2.54	VB	AT	NNS	–5.27	NNS	AT	NN
–2.72	VB	AT	NN	–5.60	JJ	AT	NN
–2.97	,	AT	NNS	–5.64	NNS	AT	NNS
–2.98	,	AT	NN	–5.95	NP	AT	NN
–3.12	VBD	AT	NN	–5.96	NP	AT	NNS
–3.28	VBD	AT	NNS	–6.26	JJ	AT	NNS
–3.41	VBG	AT	NNS	–8.21	PPSS	AT	NNS
–3.62	VBG	AT	NN	–8.90	AT	AT	NNS
–3.88	RB	AT	NN	–9.42	PPSS	AT	NN
–4.17	RB	AT	NNS				

The Proposed Method

- Conceptually, enumerate all assignments
- Score each path (product of lexical and contextual probabilities)
- Select best

Suppose *I*, *see* and *a* are each two ways ambiguous. Then there are 8 paths:

	I	see	a	bird
1.	PPSS	VB	AT	NN
2.	PPSS	VB	IN	NN
3.	PPSS	UH	AT	NN
4.	PPSS	UH	IN	NN
5.	NP	VB	AT	NN
6.	NP	VB	IN	NN
7.	NP	UH	AT	NN
8.	NP	UH	IN	NN

	.	.	I	see	a	bird	.	.	
A1	.	.	PPSS	VB	AT	NN	.	.	
context	0.99	0.20	0.07	0.07	0.23	0.25	1.00	1.00	e-4
lex	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
A2	.	.	PPSS	VB	IN	NN	.	.	
context	0.99	0.20	0.08	0.03	0.13	0.25	1.00	1.00	e-9
lex	1.00	1.00	1.00	1.00	e-4	1.00	1.00	1.00	
A3	.	.	PPSS	UH	AT	NN	.	.	
context	0.99	1.00	0.00	0.00	0.23	0.25	1.00	1.00	0
lex	1.00	1.00	1.00	e-3	1.00	1.00	1.00	1.00	
A4	.	.	PPSS	UH	IN	NN	.	.	
context	0.99	1.00	0.00	0.00	0.13	0.25	1.00	1.00	0
lex	1.00	1.00	1.00	e-3	e-4	1.00	1.00	1.00	
A5	.	.	NP	VB	AT	NN	.	.	
context	0.97	0.03	0.01	0.07	0.23	0.25	1.00	1.00	e-10
lex	1.00	1.00	e-4	1.00	1.00	1.00	1.00	1.00	
A6	.	.	NP	VB	IN	NN	.	.	
context	0.97	0.03	0.01	0.03	0.13	0.25	1.00	1.00	e-15
lex	1.00	1.00	e-4	1.00	e-4	1.00	1.00	1.00	
A7	.	.	NP	UH	AT	NN	.	.	
context	0.97	0.00	0.00	0.00	0.23	0.25	1.00	1.00	0
lex	1.00	1.00	e-4	e-3	1.00	1.00	1.00	1.00	
A8	.	.	NP	UH	IN	NN	.	.	
context	0.97	0.00	0.00	0.00	0.13	0.25	1.00	1.00	0
lex	1.00	1.00	e-4	e-3	e-4	1.00	1.00	1.00	

1. Dynamic Programming

- Conceptually, there could be k^n part of speech sequences, where *n* is the length of the input sentence, and *k* is the (worst case) lexical ambiguity.

- Fortunately, there is a linear time dynamic programming solution.

- If two paths are the same within the ngram window of 3 words, then keep the just better one.

- This way, there will be at most nk^3 paths to consider.

- *k* is small

On the Inadequacy of Ngram Models

- Ngram models were popular in the 1950s,
- until Chomsky showed that they are inadequate for many purposes.
- While they are not perfect,
- the approximation is not as bad as some others that are often made...
- For example, it is much worse to ignore lexical probabilities, as ATNs do

The Non-deterministic Non-Solution

- What's wrong with the ATN approach?

- Consider the trivial sentence: *I see a bird.*

- As we have seen, this is easy for the proposed stochastic method because every word is (almost) unambiguous:

- Prob(*I* is a pronoun) = 5837/ 5838

  Prob(*see* is a verb) = 771/ 772

  Prob(*a* is an article) = 23013/ 23019

  Prob(*bird* is a noun) = 26/ 26

- But, according to Websters, every word is ambiguous...

Word	**Parts of Speech**	
I	pronoun	noun
see	verb	noun
a	article	noun
bird	noun	verb

- One might hope that these spurious assignments could be ruled out by the parser as syntactically ill-formed.

- But if the parser is going to accept

  [NP city/ NN school/ NN committee/ NN meeting/ NN]

- then it can’t rule out

  [NP I/ NN see/ NN a/ NN bird/ NN]

- These part of speech assignments aren’t wrong; they are just extremely improbable.

Smoothing Issues

- Must do something with Zeros
- Zipf's Law: there will always be a large tail of low frequency words
- 40,000 words (80%) in the Brown Corpus have freq < 5
- If ''yawn'' appears once as a noun and once as a verb, what is the probability that it could be an adjective?

Summary of Part of Speech Discussion

- A part of speech program has been presented.
- A simple noun-phrase parser has also been implemented using similar methods.
- Performance is very encouraging.
- Lesson from speech recognition research:
  stochastic methods are often helpful when:
  - data rates are high,
  - there is plenty of training material, and
  - nothing else seems to work very well (because we don’t know what we’re doing).

- Probability vs Possibility

- Computational linguistics doesn’t like to use word frequencies, but any psycholinguist knows that they they swamp out syntactic factors

- Breadth vs Depth

Problems

Flying Planes and friends
[Time/NN] flies/VBZ like/CS [an/AT arrow/NN] ./.
[Fruit/NN] flies/VBZ like/CS [a/AT banana/NN] ./.

[Flying/VBG planes/NNS] can/MD be/BE dangerous/JJ ./.
[They/PPSS] are/BER flying/VBG [planes/NNS] ./.

Inadequate window size
[The/AT horse/NN] have/HV raced/VBN past/IN [the/AT barn/NN] ./.
[The/AT horse/NN] has/HVZ slipped/VBN ./.
[The/AT horse/NN] has/HVZ raced/VBN past/IN [the/AT barn/NN] and/CC slipped/VBD ./.

Unknown words
Do/DO [you/PPSS] know/VB [what/WDT] [a/AT xxx/NN] is/BEZ ?/.
[I/PPSS] know/VB [care/NN] if/CS [you/PPSS] xxx/VB !/.
[I/PPSS] need/MD xxx/VB ./.

Lack of word association norms, semantics,

pragramatics
[I/PPSS] like/VB to/TO work/VB ./.
[I/PPSS] went/VBD to/TO work/VB ./.
[I/PPSS] went/VBD to/IN [school/NN] ./.

Harman's Paper on Suffixing

- Main conclusion: stemming did not improve performance

- My explanation: morphologically related words tend to have modest positive correlations: $0 >> \rho >> 1$ (see figure)

- The approx that $\rho \approx 1$ is pretty terrible, and does as much harm as good.

- Don't introduce the approx for words that are common enough so that all of their forms are well-estimated.

- The approx may be justified for ''smoothing'' purposes; that is, interpolating probs for forms that are too rare to be adequately estimated.

Freq of Sg Form is Weakly Cor with Freq of Pl

Pr(noun|word)

Conclusions

- Long Queries Help
- Elaborate Queries -- Don’t Summarize them
- Natural Language Helps
  by adding bits (elaboration)
  or by removing noise (ambiguity)
- It hurts when it summarizes too much
- IR is good at finding ‘‘similar’’ stories
- Not good for understanding
  who did what to whom